Journal, Advances in Mathematical & Computational Sciences
Vol. 11 No. 3, September, 2023
www.isteams.net/mathematics-computationaljournal

Journal of Advances in Mathematical & Computational Sciences
An International Pan-African Multidisciplinary Journal of the SMART Research Group
International Centre for IT & Development (ICITD) USA

Available online at https://www.isteams.net/ mathematics-computationaljournal.info
CrossREF Member Listing - https://www.crossref.org/06members/50go-live.html

A Decision Trees-Based Model for the Classification of the Risk of Prostate Cancer

[1]Egejuru Ngozi Chidozie, [2]Balogun Jeremiah Ademola, [3]Komolafe Olufemi, [4]Egejuru Shama Chidi & [5]Idowu Peter Adebayo
[1]Department of Computer Science, Hallmark University, Ijebu-Itele, Nigeria
[2]Department of Computer Science and Mathematics, Mountain Top University, Ibafo, Ogun State, Nigeria
[3]Engineering Materials Development Institute, Akure, Nigeria
[4]Department of Computer Science, Paul University, Awka, Nigeria
[5]Department of Computer Science and Engineering, Obafemi Awolowo University,Ile-Ife, Nigeria
Corresponding Author E-Mail Address: paidowu1@yahoo.com

ABSTRACT

This paper focuses on prediction model for risk of prostate cancer. The model was formulated using the data collected which consist of information about the genomic factors and the respective risk of prostate cancer diseases from Nigeria. The WEKA software was used as the simulation environment for the development of the predictive model. The results presented in the simulation and the evaluation of the classification model developed using the C4.5 DT algorithm. The results for using the training dataset for model development showed that the values of the TP rate, FP rate and Precision had values of 1.000, 0.017 and 0.987 respectively for the tumor class while the normal class had values 0.983, 0.000 and 1.000 respectively. The study concluded that using the C4.5 decision trees algorithm a better classification model was developed within the shortest time. The study concluded that using the 6 attributes selected by the C4.5 decision trees algorithm, an effective classification model which is reliable and with a structural meaning can be developed.

Keywords: Prostate Cancer, Decision Tree, Classification Model, Algorithm, CART

Egejuru N.C., Balogun, J.A., Komolafe, O., Egejuru, S.C. & Idowu, P.A. (2023): A Decision Trees-Based Model for the Classification of the Risk of Prostate Cancer. Journal of Advances in Mathematical & Computational Science. Vol. 11 No. 3
Pp 57-76. dx.doi.org/10.22624/AIMS/MATHS/V11N3P4 Available online at www.isteams.net/mathematics-computationaljournal.

1. INTRODUCTION

Cancer prognoses are important to facilitate early cancer diagnosis, risk assessment of future events, and clinical treatment decision-making (Delen et al., 2005; Gupta et al., 2014). As a consequence,

prognostic models for disease occurrence, progression and survival are abundant for nearly all type of cancers. An accurate prediction of risks for cancer outcomes is critical for physicians and patients to make informed decisions on next steps (Katz et al., 2012).

Governments and health-care departments also rely on cancer prognostic models in planning and allocating health-care resources (Oberije et al., 2015). A typical cancer prognostic model will predict the risk of future clinical outcomes at defined time points based on certain demographic, clinical and/or genetic factors (Sesen et al., 2013). Only factors correlated with the clinical outcome of interest should be included in the model (Wang et al., 2011). These factors are called prognostic factors or risk factors, the information of which is available before the clinical endpoint of interest is observed. For example, the Prostate-Specific Antigen (PSA) and the Gleason score are known as important risk factors for prostate cancer occurrence, recurrence, and overall survival (Lowrance et al., 2010; Halabi et al., 2014).

Presently, almost all clinical practice guidelines and quality measures for process of care in cancer research are cancer stage specific or risk stratum specific in nature (Gregg et al., 2017). In order to determine whether evidence-based care is delivered to the appropriate candidate at the correct point in the course of his or her disease, one must know the cancer stage and other factors that comprise cancer risk. Although the delivery of health care services may be gleaned readily from claims data, cancer stage and risk are usually determined by examination of the medical record, a process that is often labor intensive and error prone (Luque-Baena et al., 2014). These key pieces of information are the basis for communication between researchers and clinicians; however, they remain buried deep within the electronic medical record (EMR), where they may be nevertheless accessible to automated extraction.

The increasing availability of electronic healthcare databases is enhancing opportunities for developing computer-based prediction and decision support models which can be used to improve the management of patients by healthcare professionals. An important challenge for clinical teams remains the prediction and assessment of risk, and the development of accurate approaches for diagnosing, and predicting the diagnosis and therapeutic responsiveness and outcomes (Shariat et al., 2009). The aim of predictive modeling in the context of medicine involves the development of computational models which are capable of predicting future events and/or healthcare-related outcomes for patients using contemporarily-available healthcare data (Waljee et al., 2014). These models can be based on statistical techniques or computational intelligence techniques, with the latter being a relatively new strategy.

Data mining involves the identification of unseen patterns in information stored in databases using machine learning algorithms. Data mining has a great potential to enable healthcare systems to use data more efficiently and effectively thereby reducing the likely costs associated with making decisions (Idowu et al., 2015). Data mining techniques are very useful in healthcare domain. They provide better medical services to the patients and helps to the healthcare organizations in various medical management decisions. Classification is one of the most popularly used methods of Data Mining in Healthcare sector. It divides data samples into target classes. The classification technique predicts the target class for each data points. With the help of classification approach a risk factor can be associated to patients by analyzing their patterns of diseases.

Machine learning algorithms provide a means of obtaining objective unseen patterns from evidence-based information especially in the public health care sector. These techniques have allowed for not only substantial improvements to existing clinical decision support systems, but also a platform for improved patient-centered outcomes through the development of personalized prediction models tailored to a patient's medical history and current condition (Moudani et al., 2011). To overcome this problem, medical decision support systems using data mining and machine learning is becoming more and more essential, which assists the doctors in taking correct decisions. In machine learning, feature selection is the process of selecting a subset of relevant features to construct a model by removing variables with little or no analytical value.

Feature selection is important since choosing irrelevant features would increase the time, cost, and complexity of computation and reduce the accuracy of the model (Wu et al., 2012).

There has been a number of application of machine learning algorithms on the area of cancer research including prostate cancer but most of the algorithms used have been black-boxed models which do not support the structural or mathematical representations of the relationship between predictors and targeted diseases. There is the need to develop a classification model which represents the relationship between predictors and targeted diseases using a structural hierarchical tree which shows how the predictors are related to the risk of prostate cancer. This study focuses on the application of machine learning to the discovery of unseen pattern so as to identify the most relevant of features and for the development of effective and efficient classification model for the risk of prostate cancer.

2. RELATED WORKS

Ghaheri et al. (2015), in their study presented the various applications of genetic algorithms in medicine. The study presented a review of the nature of the genetic algorithms alongside the various applications of the genetic algorithm in medicine which includes radiology, radiotherapy, oncology, pediatrics, cardiology, endocrinology, surgery, obstetrics and gynecology, pulmonology, infectious diseases, orthopedics, rehabilitation medicine, neurology, pharmacotherapy, and health care management. This review introduced the applications of the genetic algorithm in disease screening, diagnosis, treatment planning, pharmaco-vigilance, prognosis, and health care management, and enables physicians to envision possible applications of this metaheuristic method in their medical career.

Adams et al. (2015), applied genetic algorithm (GA) to variable optimization and the predictive modelling of the 5-year mortality of terminal diseases. The study examined 123 questions (variables) answered by 5,444 individuals in the National Health and Nutrition Examination Survey. The GA iterations selected the top 24 variables, including questions related to stroke, emphysema, and general health problems requiring the use of special equipment, for use in predictive modeling by various parametric and nonparametric machine learning techniques. Using these top 24 variables, gradient boosting yielded the nominally highest performance (area under curve [AUC] = 0.7654), although there were other techniques with lower but not significantly different AUC. The study shows how GA in conjunction with various machine learning techniques could be used to examine questionnaire data to predict a binary outcome.

Tan et al. (2016), worked on the application of genetic programming (GP) to the prognosis of oral cancer disease among patients. The data used for the study contained 31 cases collected from the Malaysia Oral Cancer Database and Tissue Bank System (MOCDTBS). The feature subsets that is automatically selected through GP were noted and the influences of this subset on the results of GP were recorded. In addition, a comparison between the GP performance and that of the Support Vector Machine (SVM) and logistic regression (LR) are also done in order to verify the predictive capabilities of the GP. The result showed that GP had the best performance with an average accuracy of 83.87% and area under the ROC of 0.8341 using the reduced features consisting of smoking, drinking, chewing, histological differentiation of SCC, and oncogene.

Khare and Burse (2016), applied genetic algorithm (GA) to the extraction of the most relevant features required for the classification of the risk of ovarian cancer. Data for the study was collected from the UCI Data Repository following which the genetic algorithm was applied for the extraction of the most relevant features from the initially identified features in the original dataset. The original dataset contained 216 instances with 15154 attributes defined for the two class problem of ovarian cancer.

Following the application of GA, it was observed that the features were reduces from 15154 to 22 features and were used to develop a classification model for ovarian cancer using various machine learning algorithms. The results of the study showed that the classification model developed using the reduced features selected by GA had a better performance compared to the model developed using the initially identified 15154 features.

3. METHODS

Data Collection and Pre-processing

For this study which required the development of a classification model for the risk of prostate cancer, data was collected from an online resource which was accessed from the University of Chicago Illinois (UCI) machine learning repository located online. The required dataset which contained information about breast cancer patients with risk and those without risk was downloaded from the repository as a text file which was later preprocessed into an arff file format. Following the process of data collection, the data was pre-processed using feature selection techniques in order to identify the most relevant features among the initial input features in the dataset collected.

Collection of relevant data

The dataset required for this study was downloaded from an online repository which was accessed from the University of Chicago Illinois (UCI) machine learning repository located online and retrieved from the location at https://archive.ics.uci.edu/ml/datasets/Prostate+Cancer. The dataset collected contained 136 prostate cancer data records consisting of 12601 genomic attributes which were all numeric valued. The dataset was downloaded from the repository as a comma separated variable (.csv) file format containing the attributes used to describe the data on the first row following which the data for each breast cancer patient was defined as either a recurrence or no recurrence.
The class label used to identify the recurrence of breast cancer was defined as event and was used to represent 77 records with Prostate cancer tumors while 59 records had no prostate cancer tumors. In all there were 12602 features among which 12601 were used as the input variables while one was used as the target variable for the event of breast cancer recurrence.

Pre-processing of collected data

Following the process of the identification and collection of the dataset required for the development of the classification model required for the classification of the risk of prostate cancer. The data collected was pre-process in order to identify valid and invalid values within the dataset collected in additional to the presence of missing data values also. Following the process of cleaning the data for missing and inconsistent values, the data was converted into a structured format which was required by the simulation environment. The data was converted into an attribute relation file format (.arff) which defined datasets using 3 different portions as shown in Figure 3.1.

According to the figure, the name of the file is defined by the @relation tag of the .arff file which is followed by the @attributes tag which was used to define each attribute consisting of the inputs and the output at the last @attribute line. Following the description of the name of the attribute is the definition of the values that can be given to each attribute defined with the target class which describes the risk of prostate cancer on the last line of attributes defined. Following the process of the identification and collection of the data needed for developing the predictive model, it was necessary in order to determine which set of variables are deemed more predictive for breast cancer recurrence.

Formulation of the Classification Model for the Risk of Prostate Cancer

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs. Supervised machine learning algorithms make it possible to assign a set of records (prostate cancer risk indicators) to a target classes – the diagnosis of prostate cancer (Yes or No). Supervised machine learning algorithms are Black-boxed models, thus it is not possible to give an exact description of the mathematical relationship existing among the independent variables (input variables) with respect to the target variable (output variable – risk of prostate cancer). Cost functions are used by supervised machine learning algorithms to estimate the error in prediction during the training of data for model development.

For any supervised machine learning algorithm proposed for the formulation of a predictive model, a mapping function can be used to easily express the general expression for the formulation of the predictive model for the classification of risk of prostate cancer – this is as a result that most machine learning algorithms are black-box models which use evaluators and not power series/polynomial equations. The historical dataset S which consists of the records of patients containing fields representing the set of classification factors (i number of input variables for j patients), X_{ij} alongside the respective target variable (risk of prostate cancer) represented by the variable Y_j – the risk of prostate cancer for the jth individual in the j records of data collected from the hospital selected for the study. Equation 3.1 shows the mapping function that describes the relationship between the classification factors and the target class – classification of risk of prostate cancer.

$$\varphi: X \rightarrow Y \qquad (3.1)$$
$$defined\ as: \varphi(X) = Y$$

The equation shows the relationship between the set of classification factors represented by a vector, X consisting of the values of i variables and the label Y which defines the risk of prostate cancer – Tumor and Normal for each patient as expressed in equation 3.2. Assuming the values of the set of variable for a patient is represented as $X = \{X_1, X_2, X_3, \ldots\ldots, X_i\}$ where X_i is the value of each variable, i = 1 to i; then the mapping φ used to represent the predictive model for patient performance maps the variables of each individual to their respective risk of ovarian cancer according to equation 3.2.

$$\varphi(X) = \begin{cases} Tumor \\ Normal \end{cases} \qquad (3.2)$$

The machine learning algorithms developed for the risk of prostate cancer was formulated using 2 decision trees algorithm and were benchmarked by comparison based on some metrics. Both machine learning algorithms were compared using a number of performance evaluation criteria with the best model selected from the two.

Decision trees algorithm

The formulation of the predictive model for the risk of prostate cancer was proposed using decision trees algorithm for the development of a hierarchical tree structure using a splitting criteria. The theory of decision trees has the following parts: a root node which is the starting point of the trees with branches called edges connecting successive nodes showing the flow based on the values (edge for transition) of the attribute (node) and nodes that have child nodes are called interior nodes (parent nodes). Leaf or terminal nodes are those nodes that do not have child nodes and represent a possible value of the target variable (prostate cancer risk class) given the variables represented by the path from the root node.

Rules can then be induced from the trees taking paths created from the root node all the way to their respective leaf using IF-THEN statements. The basic idea of the decision trees algorithms used was to split the given dataset into subsets by recursive partitioning of the parent nodes into child nodes based on the homogeneity of the of within–node instances or separation of between-node instances with respect to their target variables. Thus at each nodes (or identified risk factor attributes) were examined and the splitter was chosen to be the attribute such that after dividing the nodes into child nodes according to the value of the attribute variable, the target is differentiated to the best using algorithm.

Given a set X_{ij} of j number of cases, the decision trees algorithm grows an initial tree using the divide-and-conquer algorithm as follows:

- If all the cases in X_{ij} belong to the same class or X_{ij} is small, the tree is a leaf labeled with the most frequent class in X_{ij}.
- Otherwise, choose a test based on a single attribute X_i with two or more outcomes. Make this test the root of the tree with one branch for each outcome of the test, partition X_{ij} into corresponding subsets according to the outcome for each case, and apply the same procedure recursively to each subset.

The result of this algorithm also called the Hunt's algorithm is that a decision tree showing some of the attributes selected as nodes starting from the most important at the top all the way through successive nodes to the target classes at the leaves. The set of attributes that were used to construct the decision trees are the most relevant out of the initial variables identified. Also, rules were extracted from the formulated decision trees which used IF-THEN statements to combine the values of each attributes from the root nodes through successive nodes to the target class at the leaf from top to bottom.

Attribute selection criteria used by selected decision trees algorithm

As stated earlier, the decision trees algorithm requires a number of criteria for determining which attributes should be selected from splitting the dataset at each iteration of the divide-and-conquer approach used. The decision trees algorithm considered in this study were the C4.5 decision trees algorithm and the Classification and regression Trees (CART) and their respective criteria presented in the following paragraphs.

C4.5 Decision Trees (DT) algorithm

The C4.5 decision trees required the use of two criteria for selecting the most optimal attribute for splitting the dataset or decision tree. The first is called the Gain Ratio (GT), which is determined by dividing the Information Gain (IG) in equation (3.3) by the split ratio in equation (3.4). The IG is defined as the difference between the entropy (H) of an attribute H(X_i) and the entropy of the target class given the attribute X_i called H(Y|X_i) as identified in equation (3.5). Therefore, the higher the gain ratio of an attribute then the most likely its adoption as a node required for splitting the dataset.

$$IG(X_i) = H(X_i) - \sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot H(X_i) \qquad (3.3)$$

Where: $H(X_i) = -\sum_{t\epsilon T} \frac{|t,X_i|}{|X_{ij}|} \cdot \log_2 \frac{|t,X_i|}{|X_{ij}|}$

$$Split(T) = -\sum_{t\epsilon T} \frac{|t|}{|X_{ij}|} \cdot \log_2 \frac{|t|}{|X_{ij}|} \qquad (3.4)$$

and T is the set of values for a given attribute X_i.

CART Decision Trees

The Classification and Regression Trees (CART) used a merit called the Gini index to determine the attribute with the best split of dataset for identification of a tree node during the process of growing the decision tree. The Gini index is presented in equation (3.5). Let $p(C|X_i)$ denote the fraction of records that belong to a class C at a given node X_i. Therefore, the higher the Gini Index of an attribute then the most likely its adoption as a node required for splitting the dataset.

$$Gini(X_i) = 1 - \sum_{j=1}^{2} [p(j|X_i)]^2 \qquad (3.5)$$

Where: c is the number of target classes

Model Simulation Process and Environment

Following the identification of the algorithms that were needed for the formulation of the predictive model for the risk of kidney disease, the simulation of the predictive model was performed using the data collected which consisted of individuals records containing information about the genomic factors and the respective risk of prostate cancer diseases from a hospital in south-western Nigeria. The WEKA software – a suite of machine learning algorithms was used as the simulation environment for the development of the predictive model.

The dataset collected was divided into two parts: training and testing data – the training data was used to formulate the model while the test data was used to validate the model. The process of training and testing predictive model according to literature is a very difficult experience especially with the various available validation procedures. For this classification problem, it was natural to measure a classifier's performance in terms of the error rate. The classifier predicted the class of each instance; if it is correct, that is counted as a success; if not, it is an error. The error rate being the proportion of errors made over a whole set of instances, and thus measured the overall performance of the classifier. The error rate on the training data set was not likely to be a good indicator of future performance; because the classifiers were been learned from the very same training data.

In order to predict the performance of a classifier on new data, there was the need to assess the error rate of the predictive model on a dataset that played no part in the formation of the classifier. This independent dataset was called the test dataset – which was a representative sample of the underlying problem as was the training data. It was important that the test dataset was not used in any way to create the classifier since the machine learning classifiers involve two stages: one to come up with a basic structure of the predictive model and the second to optimize parameters involved in that structure.

10-fold cross validation technique

The process of leaving a part of a whole dataset as testing data while the rest is used for training the model is called the holdout method. The challenge here is the need to be able to find a good classifier by using as much of the whole historical data as possible for training; to obtain a good error estimate and use as much as possible for model testing. It is a common trend to holdout one-third of the whole historical dataset for testing and the remaining two-thirds for training. For this study the cross-validation procedure was employed, which involved dividing the whole datasets into a number of folds (or partitions) of the data. Each partition was selected for testing with the remaining k – 1 partitions used for training; the next partition was used for testing with the remaining k – 1 partitions (including the first partition used or testing) used for training until all k partitions had been selected for testing. The error rate recorded from each process was added up with the mean the mean error rate recorded. The process used in this study was the stratified 10-fold cross validation method which involves splitting the whole dataset into ten partitions.

Performance Evaluation of Model Validation Process

During the course of evaluating the predictive model, a number of metrics were used to quantify the model's performance. In order to determine these metrics, four parameters must be identified from the results of predictions made by the classifier during model testing.

Journal, Advances in Mathematical & Computational Sciences
Vol. 11 No. 3, September, 2023
www.isteams.net/mathematics-computationaljournal

These are: true positive (TP), true negative (TN), false positive (FP) and false negative (FP). True positives/negatives are correct classifications while false positives/negatives are incorrect classifications/misclassifications. These results are presented on confusion matrix – for this study the confusion matrix is a 2 x 2 owing for the 2 labels for the output class – risk of prostate cancer, namely: Tumor and Normal.

Figure 3.1 shows the diagram of the confusion matrix that was used for evaluating the performance of the decision trees algorithms developed in this study. Each cell in the 2 x 2 matrix represents the correct/incorrect classification depending on the cell referenced.

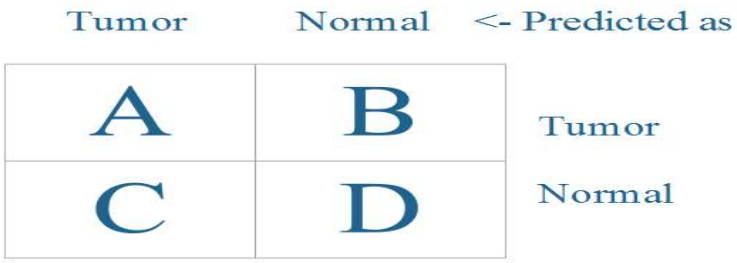

Figure 3.1: Confusion Matrix diagram for performance evaluation

The values of the cells are in turn used to estimate the performance metrics. The sum of the values of the cells across provides the number of actual cases in the training dataset while the sum of the columns provide the number of predicted cases in the training dataset. The cells located on the diagonal are the correct classifications (true positives/negatives) while other cells are the misclassifications/incorrect classifications (false positives/negatives). The performance metrics are thus defined as follows:

Sensitivity/True positive rate/Recall: is the proportion of actual cases that were correctly predicted.

$$TP\ rate_{Tumor} = \frac{A}{A+B} \qquad (3.6a)$$

$$TP\ rate_{Normal} = \frac{D}{C+D} \qquad (3.6b)$$

False Positive rate/False alarm: is the proportion of actual cases that were incorrectly predicted as another class.

$$FP\ rate_{Tumor} = \frac{C}{C+D} \qquad (3.7a)$$

$$FP\ rate_{Normal} = \frac{B}{A+B} \qquad (3.7b)$$

Precision: is the proportion of the predicted cases that were correctly predicted.

$$Precision_{Tumor} = \frac{A}{A + C} \qquad (3.8a)$$

$$Precision_{Normal} = \frac{D}{B + D} \qquad (3.8b)$$

Accuracy: is the total number of correct classifications (positive and negative)

$$Accuracy = \frac{A + D}{A + B + C + D} \qquad (3.9)$$

4. RESULTS AND DISCUSSION

Results of the Identification and Collection of Prostate Data

In order to develop the classification mode that was required for determining the risk of prostate cancer among patients, data was collected from an online repository provided by the UCI Machine Learning Repository accessed at https://archive.ics.uci.edu/ml/datasets/Prostate+Cancer. As a results of this, the data was accessed from the repository and was downloaded as an attribute relation file (.arff) format. The .arff file is the recommended file format for storing data required for the formulation of predictive models using the WEKA simulation environment. The data that was collected for this study contained 136 patients records which were defined based on the numeric values of 12601 features which were identified as genomic data. The data collected consisted of 77 patients with prostate cancer tumor and 59 patients who did not have tumors (normal cases) as shown in Table 4.1.

The data was stored in a .arff file format which consisted of 3 parts which are described in the following as shown in Figure 4.1. The first part called the relation tag stores information about the name of the file and uses the tag @relation relationName for description. The second part of the .arff file used to store the collected data stores information about the 12601 features including the target class alongside their values using the tag @attribute attributeName value. The attributeName is the name of the genomic attribute e.g. AFFX-MurIL2_at, AFFX-BioB-5 and 35052_r_at to mention a few while the value in this case is numeric since all were stored using real values.

Also, the target class which defined the risk of prostate cancer was placed as the last attribute and unlike the genomic data collected it has a nominal value hence the reason why the value is described using attributeName called Class and value called {Tumor, Normal}. The last part of the .arff file used to collect and store the data consisted of the data tag defined as @data with the record for each patient in the following lines. Each row was used to describe the values of the 12601 attributes and the target class for each patients such that the last part of each line contains the values Tumor or Normal. Following the presentation of the results of the identification and collection of prostate cancer dataset, the results of the process of model formulation using the 2 selected decision trees algorithms are presented in the following.

Journal, Advances in Mathematical & Computational Sciences
Vol. 11 No. 3, September, 2023
www.isteams.net/mathematics-computationaljournal

Table 4.1: Distribution of Target Class among Patients' Data

Target Class	Frequency	Percentage (%)
Tumor Cases	77	56.62
Normal Cases	59	43.38
Total	136	100.00

Figure 4.1: Screenshot of .arff file for Storing Data Collected

Results of Model Formulation and Simulation

The results of the study showed that the classification model for the risk of prostate cancer was formulated using 2 decision trees algorithms, namely: classification and regression trees (CART) and the C4.5 Decision Trees (DT) algorithms. The two algorithms were used to formulate the classification model on the WEKA simulation environment using 2 different training methods. The first training method involved the use of the whole dataset for training the model and then using the same dataset for testing the model developed with the correct and incorrect classifications noted.

The second training technique involved the use of the 10-fold cross validation technique which divided the dataset into 10 parts and used one part in turn as testing data while the remaining 9 parts were used for training the classification model.

Results of the CART DT algorithm

The results of the formulation and simulation of the classification model based on the CART DT for the risk of prostate cancer using the training and 10-fold cross validation techniques are presented in this section. The results of using the training technique showed that out of the original 77 tumor cases, 71 were correctly classified while 6 were misclassified as normal while out of the original 49 normal cases, 55 were correctly classified while 4 were misclassified as tumor cases. The training technique had 126 correct and 10 incorrect classifications owing for an accuracy 92.64%. The results of using the 10-fold cross validation technique showed that out of the original 77 tumor cases, 70 were correctly classified while 7 were misclassified as normal while out of the original 49 normal cases, 45 were correctly classified while 14 were misclassified as tumor cases. The training technique had 115 correct and 21 incorrect classifications owing for an accuracy 84.56%. The results of the correct and incorrect classifications made by the CART for the training and 10-fold cross validation techniques are shown in Figures 4.2 (left) and 4.2 (right) respectively.

Results of the C4.5 DT algorithm

The results of the formulation and simulation of the classification model based on the C4.5 DT for the risk of prostate cancer using the training and 10-fold cross validation techniques are presented in this section. The results of using the training technique showed that out of the original 77 tumor cases, all 77 were correctly classified while out of the original 59 normal cases, 58 were correctly classified while 1 was misclassified as a tumor case. The training technique had 135 correct and 1 incorrect classifications owing for an accuracy 99.26%. The results of using the 10-fold cross validation technique showed that out of the original 77 tumor cases, 67 were correctly classified while 10 were misclassified as normal while out of the original 59 normal cases, 41 were correctly classified while 18 were misclassified as tumor cases. The training technique had 108 correct and 28 incorrect classifications owing for an accuracy 79.41%. The results of the correct and incorrect classifications made by the CART for the training and 10-fold cross validation techniques are shown in Figures 4.3 (left) and 4.3 (right) respectively.

Tumor	Normal	<- Predicted as
71	6	Tumor
4	55	Normal

Tumor	Normal	<- Predicted as
70	7	Tumor
14	45	Normal

Figure 4.2: Results of the CART DT Algorithm

Tumor	Normal	<- Predicted as
77	0	Tumor
1	58	Normal

Tumor	Normal	<- Predicted as
67	10	Tumor
18	41	Normal

Figure 4.3: Results of the C4.5 DT Algorithm

Discussion of results of decision trees algorithm

Following the formulation and simulation of the classification model for the risk of prostate cancer using the 2 selected decision trees algorithms, 4 decision trees were grown by the algorithms such that 2 were grown for each algorithm. The results of the use of the CART DT algorithm showed that the simulation performed using the whole training dataset for model development was better than the model developed using the 10-fold cross validation technique owing the values of their correct classifications and accuracies.

The model that was developed using the whole training dataset had 3 leaf nodes with a size of 5 and it took 13.29 seconds to build while the model developed using the 10-fold cross validation technique had also 3 leaf nodes with a size of 5 but it took 10.19 seconds to build as shown in Figure 4.4. Out of the initial 12601 genomic data attributes within the dataset, the CART DT algorithm selected only 2 genomic attributes for model building namely: 37639_at and 38484_at using 3 rules.

The rules that were generated from the decision trees are as follows:

i. If (37639_at < 74.5) Then (Target Class = “Normal”);
ii. If (37639_at = 74.5) and (38484_at < 48.0) Then (Target Class = “Tumor”); and
iii. If (37639_at = 74.5) and (38484_at >= 48.0) Then (Target Class = “Normal”).

The results of the use of the C4.5 DT algorithm showed that the simulation performed using the whole training dataset for model development was also better than the model developed using the 10-fold cross validation technique owing the values of the correct classifications and accuracies. The model that was developed using the whole training dataset had 7 leaf nodes with a size of 13 and it took 3.23 seconds to build while the model developed using the 10-fold cross validation technique had also 7 leaf nodes with a size of 13 but it took 1.58 seconds to build as shown in Figure 4.5.

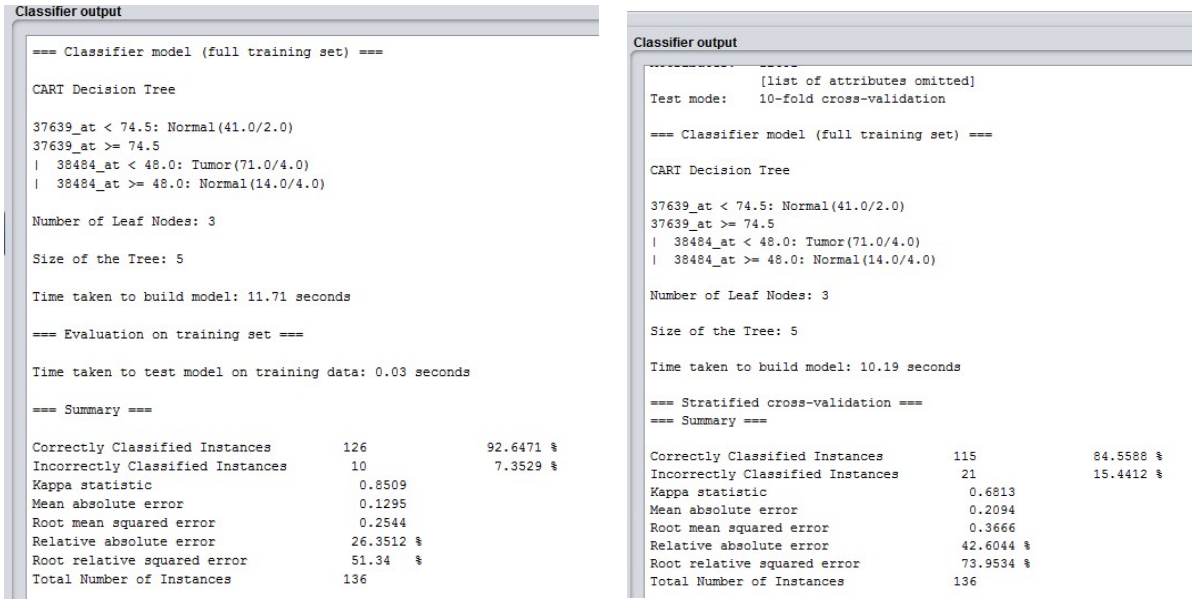

Figure 4.4: Result of Simulation for CART using Whole Training Dataset (left) and 10-Fold Cross Validation (right) Techniques

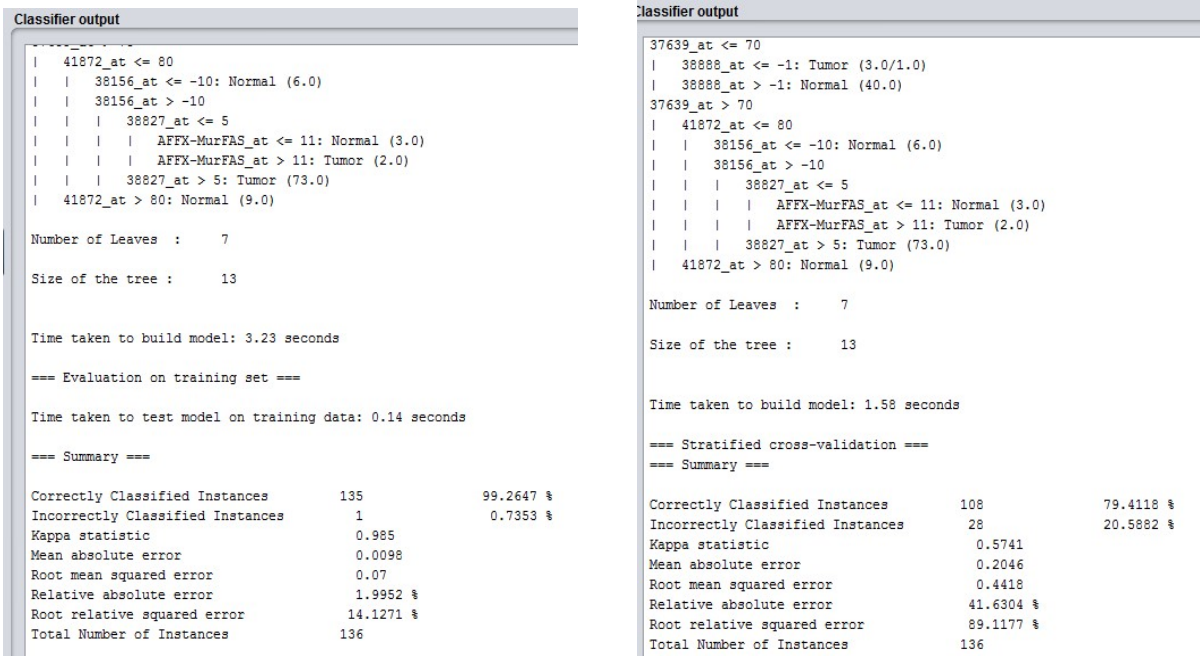

Figure 4.5: Result of Simulation for C4.5 using Whole Training Dataset (left) and 10-Fold Cross Validation (right) Techniques

Out of the initial 12601 genomic data attributes within the dataset, the CART DT algorithm selected 6 genomic attributes for model building namely: 37639_at, 38888_at, 41872_at, 38156_at, 38827_at and AFFX-MurFAS_at using 7 rules.

The decision tree grown by the C4.5 algorithm is shown in Figure 4.6.

The rules that were generated from the decision trees are as follows:

i. If (37639_at <= 70.0) and (38888_at <= -1) Then (Target Class = "Tumor");
ii. If (37639_at <= 70.0) and (38888_at > -1) Then (Target Class = "Normal");
iii. If (37639_at 70.0) and (41872_at <= 80) and (38156_at <= -10) Then (Target Class = "Normal");
iv. If (37639_at 70.0) and (41872_at <= 80) and (38156_at > -10) and (38827_at <= 5) and (AFFX-MurFAS_at <= 11) Then (Target Class = "Normal");
v. If (37639_at 70.0) and (41872_at <= 80) and (38156_at > -10) and (38827_at <= 5) and (AFFX-MurFAS_at > 11) Then (Target Class = "Tumor");
vi. If (37639_at 70.0) and (41872_at <= 80) and (38156_at > -10) and (38827_at > 5) Then (Target Class = "Tumor"); and
vii. If (37639_at 70.0) and (41872_at > 80) Then (Target Class = "Normal").

4.3 Results of Model Validation during Performance Evaluation

Following the process of the formulation and simulation of the classification models for the risk of prostate cancer from which the decision tree with the best performance was selected, the performance of each classification model was evaluated based on other relevant metrics in order to validate the model. Table 4.2 shows a summary of the results of the performance evaluation of the developed classification models from which the best was selected.

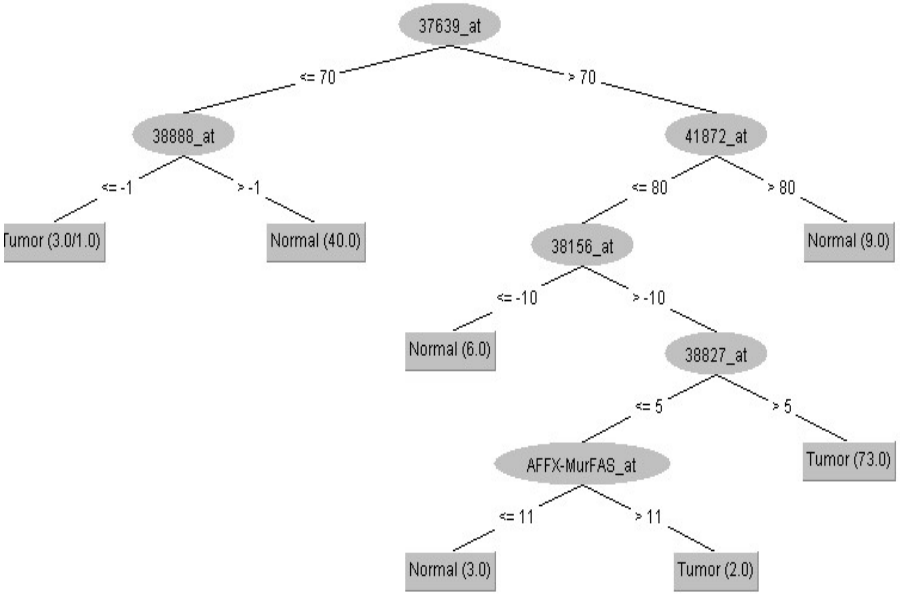

Figure 4.6: Decision Trees for Classification of Risk of Prostate Cancer

Figure 4.7 also shows the results of the evaluation of the performance of the classification models that were developed using the decision trees algorithms selected for this study. Based on the results presented in the simulation and the evaluation of the classification model developed using the CART DT algorithm, it was observed that a better model was developed using the whole training dataset for training compared to using the 10-fold cross validation technique. The results for using the training dataset for model development showed that the values of the TP rate, FP rate and Precision had values of 0.922, 0.068 and 0.947 respectively for the tumor class while the normal class had values 0.932, 0.078 and 0.902 respectively.

On an average using the whole training dataset revealed that 92.6% of actual cases were correctly classified, 7.8% of actual cases were incorrectly classified while 92.7% of predictions made by the model were correct. The results for using the 10-fold cross validation for model development showed that the values of the TP rate, FP rate and Precision had values of 0.909, 0.237 and 0.833 respectively for the tumor class while the normal class had values 0.763, 0.091 and 0.865 respectively. On an average using the whole training dataset revealed that 84.6% of actual cases were correctly classified, 1.7% of actual cases were incorrectly classified while 84.7% of predictions made by the model were correct.
Based on the results presented in the simulation and the evaluation of the classification model developed using the C4.5 DT algorithm, it was observed that a better model was developed using the whole training dataset for training compared to using the 10-fold cross validation technique. The results for using the training dataset for model development showed that the values of the TP rate, FP rate and Precision had values of 1.000, 0.017 and 0.987 respectively for the tumor class while the normal class had values 0.983, 0.000 and 1.000 respectively.

On an average using the whole training dataset revealed that 99.3% of actual cases were correctly classified, 0.1% of actual cases were incorrectly classified while 93.3% of predictions made by the model were correct. The results for using the 10-fold cross validation for model development showed that the values of the TP rate, FP rate and Precision had values of 0.870, 0.305 and 0.788 respectively for the tumor class while the normal class had values 0.695, 0.130 and 0.804 respectively. On an average using the whole training dataset revealed that 79.4% of actual cases were correctly classified, 2.3% of actual cases were incorrectly classified while 79.5% of predictions made by the model were correct.

Table 4.2: Results of the Evaluation of the Performance of Decision Trees Algorithms

Decision Trees Algorithm	Training Technique	Target Class	Correct Classifications	Accuracy (%)	True Positive (TP) rate	False Positive (FP) rate	Precision
Classification and Regression Trees (CART)	Whole Training Dataset	Tumor	126	92.65	0.922	0.068	0.947
		Normal			0.932	0.078	0.902
		Average			0.926	0.072	0.927
	10-fold Cross Validation	Tumor	115	84.56	0.909	0.237	0.833
		Normal			0.763	0.091	0.865
		Average			0.846	0.174	0.847
C4.5 Decision Trees	Whole Training Dataset	Tumor	135	99.26	1.000	0.017	0.987
		Normal			0.983	0.000	1.000
		Average			0.993	0.010	0.933
	10-fold Cross Validation	Tumor	108	79.41	0.870	0.305	0.788
		Normal			0.695	0.130	0.804
		Average			0.794	0.229	0.795

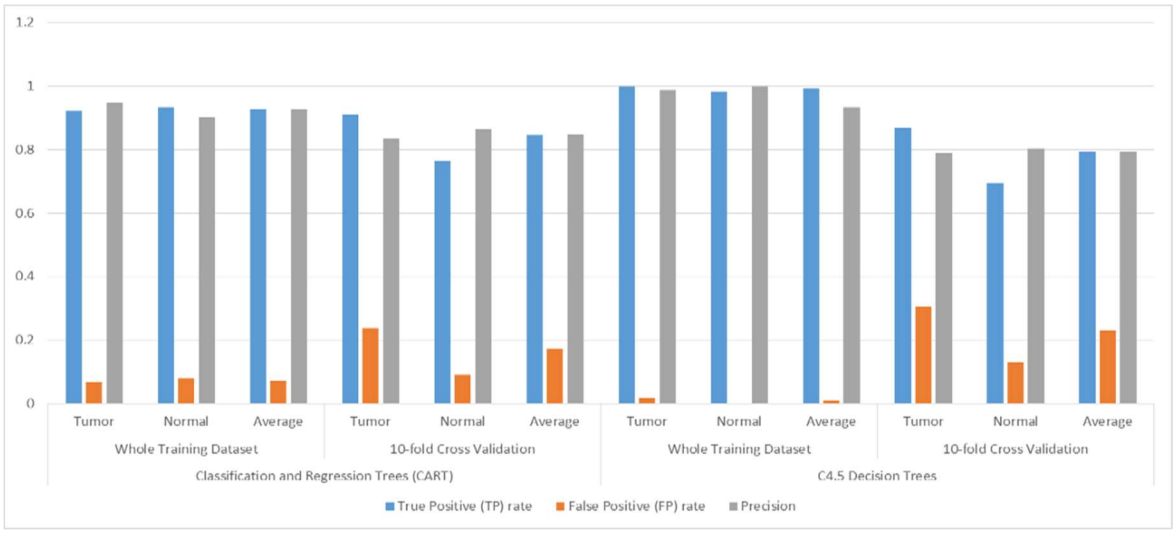

Figure 4.7: Bar Chart Plot of the Results of the Model Validation using Performance Metrics

5. CONCLUSION

The study concluded that the use of the decision trees algorithm provided a structural representation of the relationship between the genomic data identified in the collected dataset with the risk of prostate cancer using a limited number of selected genomic attributes from the original 12601 attributes. The study also concluded that although using the training dataset provided better accuracy than using the 10-fold cross validation technique however, using the 10-fold cross validation technique for classification modeling is advices as best practice which in this study provide a model with an accuracy of about 80% which is also reliable.

The study concluded that using the C4.5 decision trees algorithm a better classification model was developed within the shortest time but with a larger number of features selected compared to those selected by CART DT algorithm. The study concluded that using the 6 attributes selected by the C4.5 decision trees algorithm, an effective classification model which is reliable and with a structural meaning can be developed.

REFERENCES

Adams, L.J., Bello, G. and Dumancas, G.G. (2015). Development and Application of s Genetic Algorithm for Variable Optimization and Predictive Modelling of Five-Year Mortality Using Questionnaire Data. Journal of Bioinformatics and Biology Insights 9(3): 31 – 41.

Delen, D., Walker, G. and Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. Journal of Artificial Intelligence in Medicine 34: 113 – 127. doi:10.1016/j.artmed.2004.07.002.

Ghaheri, A., Shoar, S., Naderan, M. and Hoseini, S.S. (2016). The Applications of Genetic Algorithms in Medicine. Oman Medical Journal 30(6): 406 – 416.

Graheri, A., Shoar, S., Naderan, M. and Hoseini, S. (2015). The Applications of Genetic Algorithms to Medicine. Oman Medical Journal 30(6): 406 – 416.

Gregg, J.R., Lang, M., Wang, L.L., Resnick, M.J., Jain, S.K., Warner, J.L. & Barocas, D.A. (2017). Automating the Determination of Prostate Cancer Risk Strata from Electronic Medical Records. Report by the American Society of Clinical Oncology. JCO Clinical Cancer Informatics: 1 – 8.

Gupta, S., Tran, T. and Luo, W. (2014). Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. BMJ Open Journal 4: 1 – 13. Doi:10.1136/bmjopen-2013-004007.

Halabi, S., Lin, C.-Y. and Kelly, W.K. (2014). Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. Journal of Clinical Oncology 32: 671 – 677. doi:10.1200/JCO.2013.52.3696

Katz, M.H.G., Hu, C.-Y. and Fleming, J.B. (2012). A clinical calculator of conditional survival estimates for resected and unrested pancreatic cancer survivors. Archives of Surgery 147: 513 – 519. doi:10.1001/archsurg.2011.2281.

Khare, P. and Burse, K. (2016). Feature Selection Using Genetic Algorithm and Classification using WEKA for Ovarian Cancer. International Journal of Computer Science and Information Technologies 7(1): 194 – 196.

Lowrance, W.T., Elkin, E.B. and Jacks, L.M. (2010). Comparative effectiveness of surgical treatments for prostate cancer: a population-based analysis of postoperative outcomes. Journal of Urology 183: 1366 – 1372.

Luque-Baena, R.M., Urda, D., Subirats, J.L., Franco, L., & Jerez, J.M. (2014). Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data. Theoretical Biology and Medical Modelling 11: 1 - 18.

Mumford, C. L. & Jain, L. C. (2009). Computational Intelligence: Collaboration, Fusion and Emergence (First Edition). Springer Publishing Company, Incorporated.

Oberije, C., De Ruysscher, D. and Houben, R. (2015). A validated prediction model for overall survival from stage III non-small cell lung cancer: toward survival prediction for individual patients. International Journal of Radiation Oncology and Biological Physiology 92: 935 – 944. doi:10.1016/j.ijrobp.2015.02.048.

Sesen, M.B., Nicholson, A.E. and Banares-Alcantara, R. (2013). Bayesian networks for clinical decision support in lung cancer care. PLoS ONE 8: 1 – 12. doi:10.1371/journal.pone.0082349.

Shariat, S.F., Kattan, M.W., Vickers, A.J., Karakiewicz, P.I. & Scardino, P. T. (2009). Critical review of prostate cancer predictive tools. Journal of Future Oncology 5: 1555 - 1584.

Tan, M.S., Tan, J.W., Chang, S.-W., Yap, H.J., Kareem, S.A. and Zain, R.B. (2016). A Genetic Programming Approach to Oral Cancer Prognosis. PeerJ 4: 2482 – 24998. DOI: 10.7717/peerj.2482.

Vinterbo, S. and Ohno-Machado, L. (1999). A genetic algorithm to select variables in logistic regression: example in the domain of myocardial infarction. In Proceedings of the AMIA Symposium: 984 - 988.

Waljee, A., Higgins, P. & Singal, A. (2014). A primer on predictive models. Journal of Clinical and Translational Gastroenterology 5: 1 – 13.

Wang, S.J., Wissel, A.R. and Luh, J.Y. (2011). An interactive tool for individualized estimation of conditional survival in rectal cancer. Annals of Surgical Oncology 18: 1547 – 1552. doi: 10.1245/s10434-010-1512-3.

Wu, W.J., Lin, S.W. and Moon, W.K. (2012). Combining support vector machine with genetic algorithm to classify ultrasound breast tumor images. Computational Medical Imaging Graph 36(8): 627 - 633.